Interpretable Phonological–Semantic Dissociation Scoring for Automated svPPA Subtyping

Sarika Pasumarthy [1,*], *Minggang Li* [1,*], *Jiachen Lian* [1,**], *Chenxu Guo* [1], *Emma Xuedi Yang* [2], *Lynn Kurteff* [3], *Zoe Ezzes* [3], *Willa Keegan-Rodewald* [3], *Jet M.J. Vonk* [3], *Siddarth Ramkrishnan* [3], *Giada Antonicelli* [3], *Zachary A. Miller* [3], *Maria Luisa Gorno-Tempini* [3], *Gopala Anumanchipalli* [1]

[1] University of California, Berkeley, United States
[2] Columbia University, New York City, United States
[3] University of California, San Francisco, United States

{sarikapasumarthy, minggangli, jiachenlian, gopala}@berkeley.edu

Abstract

We introduce IPSS-PPA (Interpretable Phonological–Semantic Scoring), an interpretable, fully automated pipeline for subtyping Primary Progressive Aphasia (PPA) from picture description speech. PPA is a neurodegenerative syndrome marked by progressive language decline. Its semantic variant (svPPA) is uniquely difficult to detect because it preserves phonological production while degrading conceptual content, making it invisible to the acoustic and fluency-based biomarkers that reliably identify other variants. Task-agnostic features are mapped to ten clinician-aligned constructs, then integrated into phonological and semantic adequacy scores. Speaker diarization reduces the manual annotation bottleneck of prior work. On 254 recordings, our system achieves AUC 0.92 for control vs. svPPA and AUC 0.73 for svPPA vs. nfvPPA+lvPPA, while an LLM baseline reaches only 47% four-way accuracy. Results show that structured clinical grounding captures signal unavailable to text-only models.

Index Terms: primary progressive aphasia, picture description, automated speech analysis, clinical construct scoring, neurodegenerative disease, interpretable classification

1. Introduction

Primary Progressive Aphasia (PPA) is a neurodegenerative syndrome characterized by the progressive deterioration of language in the absence of generalized cognitive decline [1]. Gorno-Tempini et al. [2] delineate three variants with distinct neuroanatomical substrates and clinical profiles, each presenting a recognizable signature in connected speech. The *semantic* variant (svPPA) is associated with degeneration of the ventral temporal stream [3], resulting in profound loss of conceptual knowledge while leaving phonological production and speech fluency largely preserved. The *nonfluent* variant (nfvPPA) implicates left posterior frontal and insular cortex, yielding effortful, apraxic speech with marked agrammatism and syntactic simplification. The *logopenic* variant (lvPPA) arises from left temporoparietal atrophy, producing word-finding pauses and phonological working-memory deficits in the context of otherwise preserved grammar [2]. These profiles are clinically and prognostically distinct, and reliable automated identification from naturalistic speech would support both large-scale screening and longitudinal disease monitoring.

Early work explored automatic classification of PPA variants using hand-crafted acoustic, prosodic, and linguistic features [4, 5, 6, 7], and more recent approaches have leveraged end-to-end models and dysfluency-based intermediate representations [8, 9, 10, 11]. While these methods reliably characterise nfvPPA and lvPPA, where motor speech errors and dysfluency are prominent, *characterising svPPA and distinguishing it from the other variants remains the central unsolved challenge*: svPPA primarily manifests as a semantic deficit rather than a disruption of fluency. Speech remains fluent and grammatically well-formed, yet semantic content is markedly impoverished. Patients substitute underspecified referential expressions for specific lexical items, producing utterances such as *"she's holding that kite thing... with the string bits"* instead of precise descriptions. Under the dual-stream model of speech processing [3], svPPA selectively impairs the ventral (lexical-semantic) stream while sparing the dorsal (articulatory-phonological) stream, yielding a dissociation between preserved phonological fluency and degraded semantic content that neither fluency-based nor dysfluency-based metrics can capture. Concretely, acoustic-prosodic approaches achieve only 66% svPPA accuracy [5] compared to 82% for nfvPPA, and the prior state-of-the-art system [9] does not report svPPA vs. nfvPPA+lvPPA discrimination at all.

This semantic-phonological dissociation motivates direct evaluation of semantic content. Picture description tasks, in particular the Western Aphasia Battery (WAB) Picnic scene [12] and the Boston Diagnostic Aphasia Examination Cookie Theft picture, are standard clinical instruments for eliciting connected speech. They provide a normative visual referent against which the completeness and accuracy of descriptions can be objectively assessed. The WAB Picnic task is particularly amenable to automated analysis: its well-defined inventory of entities, spatial relations, and action relations admits formal representation as a scene graph and serves as a reference for stimulus-grounded semantic scoring.

To address these limitations, we introduce IPSS-PPA, an end-to-end automated pipeline for PPA subtype classification. On 254 WAB Picnic recordings, IPSS-PPA achieves AUC 0.92 for control vs. svPPA and AUC 0.734 ($p < 10^{-5}$, Cliff's $\delta = 0.469$) for svPPA vs. nfvPPA+lvPPA. Three-way subtype accuracy reaches 60.3%, comparable to the prior state of the art (74.0% on a smaller, manually annotated cohort of $n = 59$) while using 32 features versus 363 and eliminating manual di-

[*] These authors contributed equally.
[**] indicates the corresponding author.

arization entirely. A transcript-only LLM baseline achieves only 50% on svPPA vs. nfv+lv (AUC 0.500), confirming that structured phonological and semantic grounding captures signal unavailable to text-only models. Our contributions are:

- **First automated result for svPPA vs. nfvPPA+lvPPA discrimination.** To our knowledge, no prior automated system reports classification performance on this clinically consequential distinction. On 229 recordings from a longitudinal clinical cohort, our system achieves AUC 0.734 ($p < 10^{-5}$, Cliff's $\delta = 0.469$), with group-level construct scores showing svPPA's highest anomia (0.58) and lowest scene-graph coverage ($S_{\text{sem}}^{\text{graph}}$: HC 0.27 vs. svPPA 0.12) emerging without any supervised optimisation on the classification labels.
- **Semi-automated preprocessing.** Automatic speaker diarization (pyannote [13]) flags segments likely to contain mixed speakers, reducing the manual annotation burden that Peters et al. identified as a principal barrier to scalable deployment. Of 254 recordings, 174 (68.5%) pass through automatically without any manual intervention; because pyannote errors concentrate in clinician-turn and overlap regions (16.9% of total speech), hands-on review time is reduced by an estimated 68–83% relative to fully manual diarization.
- **Stimulus-agnostic semantic scoring.** The semantic scoring component is parameterized by an exchangeable JSON scene graph encoding the normative content of the eliciting stimulus. Extending the pipeline to a new picture description task (e.g., Cookie Theft or other BDAE stimuli) requires only replacing this graph file, without model retraining or feature re-engineering.
- **Structured and interpretable feature design.** Rather than relying on large undifferentiated feature sets, the pipeline computes two theoretically motivated adequacy scores, phonological adequacy (S_{ph}) and semantic adequacy (S_{sem}), together with a hierarchical layer of ten population-normed clinical constructs (including anomia, agrammatism, and empty speech) co-designed with speech-language pathologists (SLPs) to align with established QAB diagnostic categories. The model produces clinician-readable evidence chains linking predictions to named clinical features, enabling direct inspection without post-hoc explanation (see Figure 4).

2. Related Work

Hand-crafted Feature Approaches. Early automated PPA characterization relied on task-specific hand-crafted features. Acoustic and prosodic measures, including pause rate and fundamental frequency range [4], reliably capture the articulatory breakdowns characteristic of nfvPPA (82% accuracy [5]) but leave svPPA largely undetected (66%), since svPPA speech is fluent and prosodically intact. Linguistic feature approaches extended this line of work: Fraser et al. [6] distinguished svPPA from nfvPPA at 79% accuracy using syntactic and semantic NLP features, while Zimmerer et al. [14] achieved 90% control-vs.-PPA separation but only 59.4% diagnostic subgroup accuracy using word-frequency and collocation measures, underscoring the persistent difficulty of fine-grained subtype discrimination. Themistocleous et al. [7] combined acoustic and morphosyntactic features in a deep neural network to reach 80% three-way accuracy, correctly identifying 90% of nfvPPA and 95% of lvPPA cases, yet svPPA remained the hardest variant to isolate. Across these approaches, features are hand-crafted for specific stimuli or corpora and do not generalise readily to new tasks or recording conditions.

End-to-End and LLM-Based Approaches. More recent work has moved toward end-to-end and large-model methods. Rezaii et al. [8] applied unsupervised LLM clustering to 78 PPA patients' speech, achieving 88.5% agreement with clinical diagnoses, while Phan et al. [15] established a transcript-only LLM upper bound for picture-description classification. On the multimodal side, Favaro et al. [16] automated content-unit (CU) scoring for Alzheimer's detection, and Ambadi et al. [17] proposed spatio-semantic graph representations for neurodegenerative speech; neither has been validated for PPA variant discrimination. The current state of the art, Peters et al. [9], integrates acoustic features, ASR-derived linguistics, CU scoring, and spatio-semantic graphs to reach 97% binary and 74% three-way accuracy on Cookie Theft recordings. Despite these strong results, the system requires manual diarization and human-in-the-loop preprocessing at inference time, and critically does not report performance on the svPPA vs. nfvPPA+lvPPA binary discrimination. Across LLM and end-to-end methods more broadly, predictions are not decomposable into clinically named features, limiting their utility for diagnosis support and prospective validation.

Dysfluency Modeling as an Intermediate Representation. A parallel line of work has framed dysfluency as an interpretable intermediate behavioural feature, systematically defining word- and phoneme-level insertions, deletions, replacements, prolongations, pauses, and repetitions [11, 18, 19, 20]. Automated pipelines using expert-designed simulators [21, 22], transcribers [23, 24], and aligners [25] demonstrate effective screening of lvPPA and nfvPPA with strong alignment to clinician judgments [10, 26]. However, dysfluency modeling primarily captures motor speech errors and pays limited attention to semantic content, leaving a gap between proxy dysfluency signals and the clinically meaningful semantic features needed to detect svPPA. IPSS-PPA directly addresses this gap by grounding semantic scoring in a stimulus-derived scene graph and mapping features to clinical constructs that include anomia and empty speech alongside the fluency-based constructs that prior work has handled well.

3. Data

All recordings were drawn from a longitudinal clinical PPA cohort. Participants were diagnosed per Gorno-Tempini 2011 criteria [2] by board-certified neurologists. The dataset comprises 254 WAB Picnic picture-description recordings across four diagnostic groups (Table 1); multiple sessions per participant were included, with one recording per session.

The WAB Picnic scene includes 21 entities, 23 relations, 6 attributes, and 12 clinical content units per the WAB manual [12]. Figure 2 illustrates entity-level naming rates across the scene for healthy controls (HC) and svPPA participants: green bounding boxes ($\geq$50% of participants named the entity), orange (25–50%), and red ($<$25%). svPPA patients show systematically lower naming rates across salient scene entities, consistent with impaired lexical-semantic access in the presence of preserved phonological fluency. This annotated scene graph motivates the entity- and relation-level scoring in $S_{\text{sem}}^{\text{graph}}$ (Section 3.4).

3.1. Data preprocessing

Audio is naturalistic clinical speech with heterogeneous recording conditions. We apply light denoising and level normal-

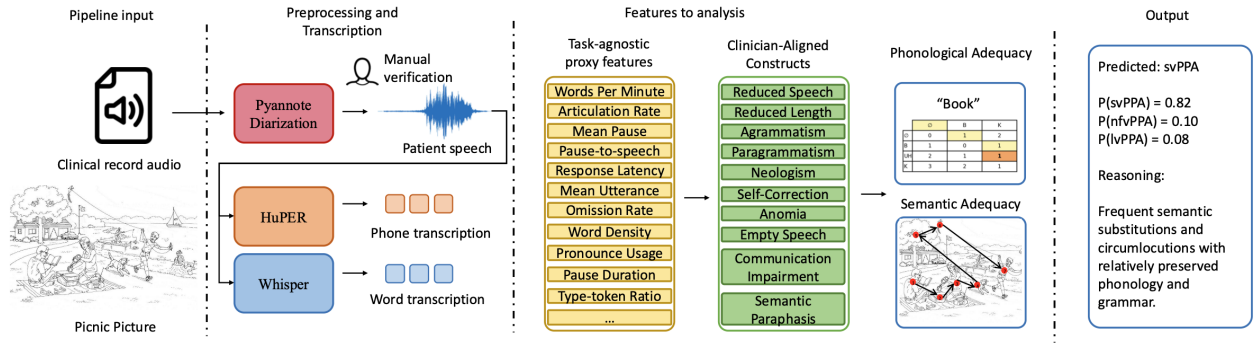

Figure 1: *Overview of the IPSS-PPA pipeline. Raw picture-description audio is semi-automatically diarized, processed into proxy features, aggregated into clinician-aligned constructs, and combined with semantic graph scoring to produce interpretable phonological and semantic adequacy scores for classification.*

Table 1: *Dataset composition and duration summary for WAB Picnic recordings.*

Group	Recordings	Label	Avg len (s)	Var (s^2)	Total len (h)
Healthy controls	25	0	83.89	1432.48	0.58
svPPA	42	1	90.00	2347.65	1.05
nfvPPA	78	2	104.29	2638.63	2.26
lvPPA	109	3	113.60	3633.22	3.44
Total	**254**		103.92	3010.43	7.33

ization before diarization to stabilize downstream ASR and phoneme extraction in low-SNR segments.

The preprocessing workflow is explicitly semi-automated: pyannote [13] is used as a first-pass speaker segmentation and overlap detector; flagged mixed-speaker segments are then corrected by trained annotators, while high-confidence single-speaker segments are accepted automatically. This design preserves throughput while maintaining clinically reliable patient-only speech segments.

The gold scene graph used for semantic scoring was developed in collaboration with clinician partners and SLPs, encoded in JSON format, and provided with the supplementary materials (anonymised). This is, to our knowledge, the *first automated scoring system validated on the WAB Picnic stimulus*.

3.2. Known Pyannote Limitations

Pyannote is benchmarked primarily on meeting, broadcast, and telephone corpora; its pretrained models are not fine-tuned to clinical interview speech, which exhibits distinctive acoustic conditions: variable recording hardware, low-SNR environments, dysarthric or slow patient speech, and frequent short clinician backchannels and prompts [13]. The principal failure modes observed in our corpus are: (1) *short backchannel mis-attribution*: brief clinician acknowledgements ("mm-hm", "okay") are occasionally assigned to the patient speaker, introducing spurious tokens into the transcript; (2) *overlap under-detection*: simultaneous speech at turn boundaries, common when a clinician provides a prompt before the patient has fully finished, may be assigned to a single speaker rather than flagged; and (3) *speaker confusion on pathological voices*: severely dysarthric or hypophonic speech (characteristic of advanced nfvPPA) can produce speaker embeddings that cluster inconsistently across a recording, leading to spurious speaker-change events. These failure modes are concentrated at segment boundaries and in short (<1 s) turns, making them identifiable and correctable by a trained annotator without requiring a full manual pass.

Table 2: *Semi-automation efficiency: two-experiment summary.*

Experiment	Metric	Est. reduction
Recording-level	174/254 auto pass-through	68.5%
Duration-based	16.9% clinician fraction	∼83%

3.3. Semi-automation Efficiency

We quantified efficiency gains along two complementary axes. *(1) Recording-level pass-through:* of the 254 recordings, 174 (68.5%) were accepted automatically without any manual intervention; the remaining 80 (31.5%) contained flagged mixed-speaker or low-confidence regions and required annotator correction, yielding a 68.5% reduction in the number of recordings requiring human diarization effort. *(2) Duration-based effort:* parsing all diarized output files and summing segment durations by speaker, participant speech totalled 12,879 s and clinician speech totalled 2,613 s (15,492 s combined). Because pyannote errors are concentrated in clinician-turn and overlap regions, hands-on review scales with the clinician fraction (16.9%), implying an approximately 83% reduction in hands-on diarization time relative to fully manual labelling of all speech. Together, these estimates suggest that the semi-automated workflow reduces total annotation burden by 68–83% depending on the metric, consistent with the diarization bottleneck identified in prior work [9].

4. Methods

Our pipeline converts raw picture-description audio into clinically grounded features in two stages, followed by classifica-

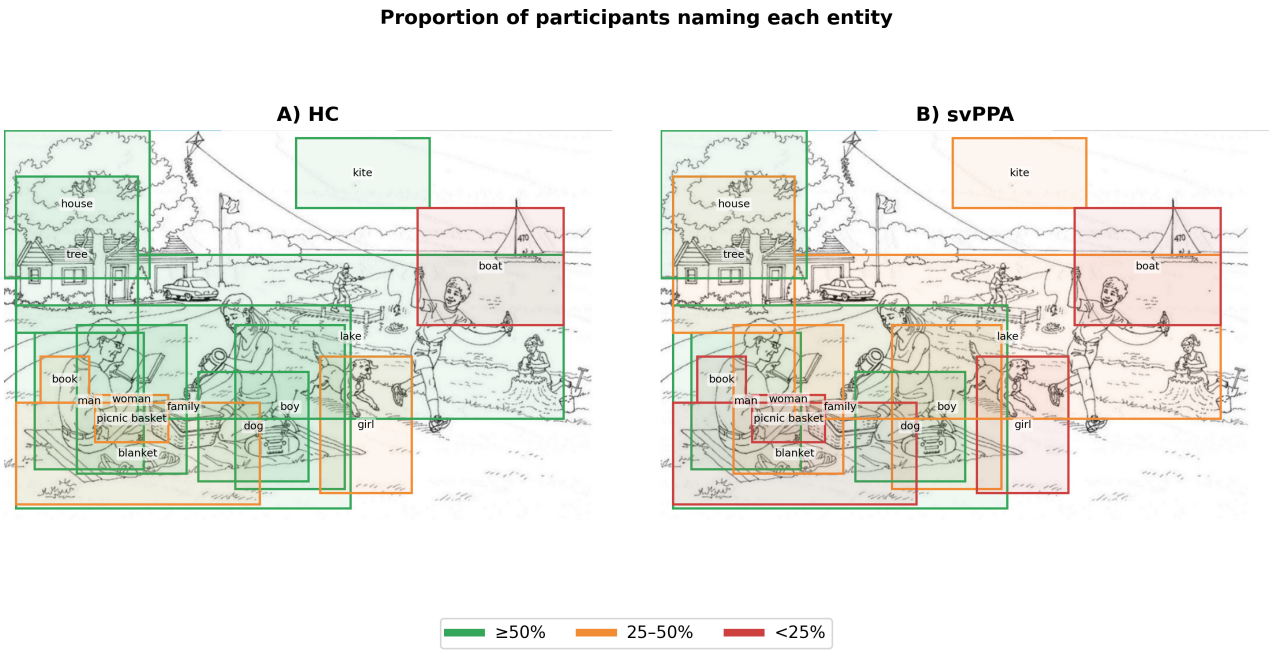

Figure 2: *Proportion of participants naming each WAB Picnic scene entity, for healthy controls (HC, panel A) and svPPA (panel B). Bounding box colour indicates naming rate: green ≥50%, orange 25–50%, red <25%. svPPA participants show reduced naming rates across entities relative to HC, reflecting degraded lexical-semantic access while phonological fluency is preserved. Entity coverage from this scene grounds the* $S_{\text{sem}}^{\text{graph}}$ *scoring component.*

tion. Layer 1 extracts task-agnostic proxy features from transcripts and timing, and Layer 2 aggregates these into clinician-aligned constructs that feed the final classifiers. Participant speech is first isolated from clinician prompts using pyannote.audio (v3.1) [13], removing the manual diarization bottleneck of prior work [9]. Two parallel branches then operate on the retained participant segments: Whisper large-v2 [27] produces a word-level transcript with timestamps, and HuPER [24] produces a lexicon-free phoneme sequence directly from the audio signal; HuPER is a state-of-the-art phoneme model. Deriving phonemes acoustically rather than from the ASR hypothesis is critical: it ensures S_{ph} reflects articulatory production quality independently of lexical substitution errors, preserving the separation between the phonological and semantic branches.

4.1. Layer 1: proxy features

Layer 1 extracts 20 task-agnostic proxy measures from the word-level transcript and pyannote timing: *words per minute, articulation rate (phones/sec), mean pause duration, pause-to-speech ratio, initial response latency, mean length of utterance, median utterance length, clauses per utterance, abandoned utterance rate, utterance count, open/closed-class ratio, content word density, type–token ratio, pronoun usage ratio, function-word omission rate, morphological error rate, dependency parse completeness, word-finding pause duration, filled-pause frequency, semantic distance to expected content.* This Layer 1 list is not intended to be exhaustive. We developed it iteratively with speech-language pathologists (SLPs), identifying proxies one by one to capture hallmark cues of different PPA variants while ensuring each proxy is directly measurable from existing automated tools (diarization, transcript, timing, and phoneme streams) without human reference annotations. spaCy (en_core_web_sm) [28] provides POS tagging and dependency parsing for morphosyntactic features. Each proxy is z-scored against cohort statistics estimated from training folds of each cross-validation split, so that feature values are interpreted relative to the population rather than in absolute terms.

4.2. Layer 2: clinical constructs

Layer 2 aggregates proxy z-scores into ten clinician-aligned constructs derived from the Quick Aphasia Battery (QAB) framework, which is designed for "reliable and multidimensional assessment of language function" [29]. In collaboration with SLPs, we mapped these QAB-aligned constructs to specific Layer 1 proxy features and weights. The ten constructs are: *reduced speech rate, reduced length/complexity, agrammatism, paragrammatism, anomia, empty speech, semantic paraphasias, phonemic paraphasias/neologisms, self-correction, and overall communication impairment.* Each construct is a weighted combination of its associated Layer 1 z-scores, passed through a bounded nonlinearity to produce a value in $[0, 1]$:

$$\text{L2}_c = \frac{\tanh(\bar{z}_c) + 1}{2} \tag{1}$$

Construct definitions and proxy weightings were specified in collaboration with SLPs to reflect the clinical categories of the WAB and QAB [29], ensuring the feature space mirrors established diagnostic reasoning. These Layer 1→Layer 2 mappings are manual and theory-driven rather than learned end-to-end. We expect to expand and refine the Layer 1 proxy inventory as additional labeled samples become available. A Layer 2 communication adequacy score summarises across constructs: $S_{\text{sem}}^{\text{L2}} = 1 - \text{mean}(\text{L2}_1, \ldots, \text{L2}_{10})$.

4.3. Phonological adequacy (S_{ph})

Expected phonemes are generated from the Whisper transcript [27] using a CMU-based G2P lookup with a character-level fallback for out-of-vocabulary items [30]. Observed phonemes are read from the HuPER stream after removing silence tokens [24]. Phonological adequacy is defined as

$$S_{\mathrm{ph}} = 1 - \frac{\mathrm{Levenshtein}(\hat{P},\, P^*)}{\max(|\hat{P}|,\, |P^*|)} \tag{2}$$

where $\hat{P}$ and P^* are the observed and canonical phoneme sequences respectively. $S_{\mathrm{ph}} \in [0, 1]$, with higher values indicating more intact phonological production.

4.4. Semantic adequacy (S_{sem})

Semantic adequacy combines a stimulus-grounded scene-graph score with the task-agnostic Layer 2 adequacy score:

$$S_{\mathrm{sem}} = \alpha\, S_{\mathrm{sem}}^{\mathrm{graph}} + (1-\alpha)\, S_{\mathrm{sem}}^{\mathrm{L2}}, \quad \alpha = 0.5 \tag{3}$$

The graph score matches the Whisper transcript against a normative WAB Picnic scene graph comprising 21 entities, 23 directed relations, and 12 clinician-defined content units [12]. Entity mentions are detected via canonical entity maps with synonym sets; relations are extracted using directional rules that respect word order, so that *boy holding string* and *string holding boy* yield distinct and differently scored edges. This directionality penalizes semantically anomalous constructions characteristic of svPPA output. Scene attributes (e.g., *kite–flying*) are extracted separately. Graph-based semantic adequacy weights relational content unit coverage more heavily than entity recall, following Nicholas & Brookshire [12]:

$$S_{\mathrm{sem}}^{\mathrm{graph}} = 0.4\, \frac{|\text{matched entities}|}{21} + 0.6\, \frac{|\text{matched CUs}|}{12} \tag{4}$$

Unlike the spatio-semantic graph of Peters et al. [9], which encodes the temporal order of content-unit mentions weighted by image coordinates, our graph encodes semantic relations between entities. Both approaches are deterministic and require no neural parser; the distinction lies in what the graph edges represent: scene navigation versus semantic content coverage.

5. Experiments

5.1. Evaluation protocol

Unless otherwise specified, we report mean $\pm$ standard deviation over 10 runs (seeds 0–9), each using stratified 5-fold cross-validation. For binary tasks, we report accuracy, F1, and ROC-AUC. For imbalanced settings, we additionally emphasize class-sensitive metrics (balanced accuracy, macro-F1, PR-AUC) and confusion matrices.

5.2. Four experiments: design and reasoning

We structure evaluation as four progressively harder experiments designed to test whether clinically grounded phonological and semantic signals support accurate and robust PPA subtype modeling.

5.2.1. Experiment 1: Control vs. svPPA

Goal. Test whether core adequacy scores distinguish svPPA from healthy controls.
Reasoning. svPPA is characterized by degraded semantic retrieval with relatively preserved phonological production. In contrast, healthy controls are expected to remain stronger and more balanced across both channels.
Design. Binary classification (Control vs. svPPA; $n = 67$) with ablations: (1) S_{sem} only, and (2) $S_{\mathrm{ph}} + S_{\mathrm{sem}}$.
Interpretation criterion. Improvement after adding S_{ph} supports the value of jointly modeling phonological adequacy and semantic deficit.

5.2.2. Experiment 2: svPPA vs. (nfvPPA + lvPPA) (subtype discrimination within PPA)

Goal. Evaluate whether the feature framework distinguishes svPPA from other PPA variants, not only from healthy controls.
Reasoning. All PPA subtypes are language-impaired, but in different ways: svPPA is semantically impaired with relatively preserved phonology; nfvPPA is typically more phonologically/syntactically impaired; lvPPA often shows phonological working-memory limitations. This overlap makes subtype separation harder and tests whether our feature set remains informative under clinically realistic conditions.
Design. Binary classification (svPPA vs. nfvPPA+lvPPA; $n = 229$) using feature groups: phonological/semantic adequacy scores, acoustic features, scene-graph semantic features, and Layer-2 clinical constructs. We evaluate logistic regression (LR) and random forest (RF).
Interpretation criterion. If models with Layer-2 constructs and multimodal features outperform simpler feature sets, this indicates complementary subtype-discriminative value from clinician-grounded abstractions and acoustic-semantic evidence.

5.2.3. Experiment 3: Balanced svPPA vs. nfvPPA+lvPPA (is performance robust to class imbalance?)

Goal. Verify that subtype discrimination is not an artifact of prevalence imbalance.
Reasoning. The full cohort is skewed (svPPA: 42; nfvPPA+lvPPA: 187). A model can obtain inflated accuracy by favoring the majority class, so an explicit balanced test is required to validate minority-class recognition.
Design. Balanced binary evaluation with repeated subsampling (25 svPPA vs. 25 nfvPPA+lvPPA per run), reporting balanced accuracy, macro-F1, PR-AUC, and confusion matrices.
Interpretation criterion. Stable performance under balancing suggests the model captures genuine svPPA-specific linguistic structure rather than class-frequency shortcuts.

5.2.4. Experiment 4: Three-way svPPA vs. nfvPPA vs. lvPPA (can the framework support full subtype diagnosis?)

Goal. Test whether the feature framework generalizes from binary decisions to complete PPA subtype classification.
Reasoning. Binary settings can hide inter-subtype confusion. Clinically useful decision support requires simultaneous differentiation of svPPA, nfvPPA, and lvPPA.
Design. Three-way classification on the full PPA cohort ($n = 229$), with macro-F1, per-class F1, and confusion-matrix analysis.
Interpretation criterion. Confusion structure indicates which distinctions remain challenging (e.g., svPPA–lvPPA overlap

Table 3: *Control vs. svPPA (top models; 10 runs).*

Feature	Acc.	F1	AUC
$S_{ph} + S_{sem}$	0.82 ± 0.01	0.85 ± 0.01	$\mathbf{0.92 \pm 0.01}$
$S_{ph} + S_{sem} + \Delta$	$\mathbf{0.82 \pm 0.01}$	$\mathbf{0.85 \pm 0.01}$	$\mathbf{0.92 \pm 0.01}$
Dissociation+extras	0.82 ± 0.01	0.85 ± 0.01	0.91 ± 0.01

in semantic symptoms) and which are better captured (e.g., nfvPPA motor-speech related profile).

6. Results

6.1. Quantitative performance

Effect sizes are reported as Cliff's δ: $|\delta| < 0.147$ negligible, 0.147–0.330 small, 0.330–0.474 medium, ≥ 0.474 large. Group differences use two-sided Mann-Whitney U (scipy.stats.mannwhitneyu). All classification results are reported as mean $\pm$ standard deviation over 10 runs (seeds 0–9), with stratified 5-fold CV re-sampled per run. This section reports results for the four experimental settings defined in Section 5.2.

6.1.1. Experiment 1: Control vs. svPPA ($n = 67$)

Table 3 reports the top-performing configurations for control vs. svPPA.

6.1.2. Experiments 2–4: svPPA vs. nfvPPA+lvPPA and 3-way PPA classification

Table 4 consolidates results across Experiments 2–4. On the real imbalanced cohort ($n = 229$), the best model (Full+L2+G, RF) achieves Acc 0.82 and AUC 0.75. The balanced 25/25 subsample confirms this result holds under equal class sizes (Bal-Acc 0.67, PR-AUC 0.72; confusion: TN 16.8, FP 8.2, FN 8.4, TP 16.6). Three-way classification (sv/nfv/lv, $n = 229$) yields macro-F1 of 0.560 with per-class F1 of 0.489 (sv), 0.641 (nfv), and 0.549 (lv).

6.2. Longitudinal analysis

To characterize longitudinal progression by subtype, we compare mean percent change from each participant's first visit for svPPA (n=10) versus pooled lvPPA+nfvPPA (n=37) over 0, 1, and 2 years. We track phonological adequacy (S_{ph}), semantic adequacy (S_{sem}), and $S_{\text{overall}} = (S_{\text{ph}} + S_{\text{sem}})/2$.

Figure 3 shows a subtype-specific dissociation consistent with the svPPA hallmark profile: svPPA exhibits a larger drop in semantic score than in phonological score, whereas lvPPA+nfvPPA shows the opposite tendency, with steeper decline in phonological score and comparatively smaller semantic change. This longitudinal separation between semantic and phonological channels supports the clinical motivation for modeling both channels explicitly in IPSS-PPA.

6.3. Baseline comparison

6.3.1. LLM transcript-only baseline

We evaluated qwen3.5-plus, a current state-of-the-art LLM, on Whisper transcripts with binary tasks and few-shot prompting (task descriptions and prompts available in the project repository). For each task, we randomly sampled 25 examples per class (50 total).

These baselines show that transcript-only LLMs underperform structured acoustic and clinically grounded features, motivating IPSS-PPA's multimodal design.

6.3.2. Comparison with prior state of the art

Table 6 summarizes LFTK vs. IPSS-PPA on identical data splits.

Direct accuracy comparisons are not strictly meaningful because the datasets and stimuli differ (Cookie Theft, $n = 59$ vs. WAB Picnic, $n = 254$). Our gains lie in scalability and interpretability: IPSS-PPA uses far fewer, clinically grounded features with transparent evidence chains, and is evaluated on a larger cohort, supporting better transferability and generalizability.

Our system achieves lower binary accuracy (0.851 vs. 0.970), attributable partly to full automation (manual diarization and preprocessing likely improve raw feature quality) and partly to stimulus differences (WAB Picnic vs. Cookie Theft). On three-way subtype accuracy, our result (82.1%) is comparable-to-better than the reported reference value (74.0%) while requiring no manual annotation and using over an order of magnitude fewer features.

6.4. Interpretability validation

6.4.1. Group-level construct profiles

Figure 4 visualises mean adequacy scores and Layer 2 clinical construct scores by diagnostic group. The top panel shows that healthy controls achieve the highest S_{ph} and S_{sem} scores, with svPPA exhibiting the sharpest drop in scene-graph coverage ($S_{\text{sem}}^{\text{graph}}$: HC 0.27 vs. svPPA 0.12), consistent with impaired semantic content generation in the presence of preserved phonological fluency. The bottom panel shows that Layer 2 construct scores are clinically coherent across groups: svPPA shows the highest anomia (0.58) among PPA subtypes, while nfvPPA shows disproportionately elevated agrammatism and the highest reduced speech rate (0.56), patterns that directly mirror the Gorno-Tempini consensus criteria [2]. These group-level patterns emerge from construct weights specified through clinical consultation, without any supervised optimisation on the classification labels, providing evidence that the Layer 2 framework captures clinically valid structure in the data independently of the classifier.

A representative svPPA case (sub-10119) is classified as svPPA with an estimated 55.2% semantic impairment score. The supporting evidence chain, expressed directly in clinical construct scores: anomia 82.8%, semantic paraphasias 54.2%, empty speech 44.1%, semantic adequacy 44.8%, sitting well above the svPPA group means visible in Figure 4. Because each construct is a named, weighted combination of proxy features, the prediction can be communicated to a clinician without reference to any opaque logit or embedding, and can be directly verified against observed patient behaviour. This feature-level transparency is a core architectural property of IPSS-PPA, not a post-hoc explanation.

6.4.2. Failure-case analysis

We select one correctly classified and one misclassified example from the control vs. svPPA task ($S_{ph}+S_{sem}$ logistic regression, 5-fold CV) to test whether the evidence chain remains clinically coherent when predictions succeed and when they fail.

Case A: correctly classified control (sub-1619, ses-20171206; true: control, predicted: control). The participant

Table 4: *Task 2 results across evaluation conditions (top models; 10 runs). Real: imbalanced cohort ($n = 229$); F1 not reported due to class imbalance. Balanced: 25 svPPA vs. 25 nfv+lv per run. 3-way: sv/nfv/lv ($n = 229$).*

Condition	Model	Acc.	AUC	Macro-F1
Real ($n = 229$)	Full+L2 (RF)	0.81 ± 0.01	0.74 ± 0.01	—
	Acoustic+graph (RF)	0.78 ± 0.02	0.65 ± 0.02	—
	Full+L2+G (RF)	$\mathbf{0.82 \pm 0.01}$	$\mathbf{0.75 \pm 0.01}$	—
Balanced (25/25)	Full+L2 (RF)	$\mathbf{0.67 \pm 0.06}$	0.72 ± 0.10	$\mathbf{0.67 \pm 0.06}$
	Full+L2+G (LR)	0.66 ± 0.03	0.70 ± 0.09	0.66 ± 0.03
	Full+L2+G (RF)	0.66 ± 0.05	0.71 ± 0.08	0.66 ± 0.05
3-way (sv/nfv/lv)	RF (best Acc.)	$\mathbf{0.60 \pm 0.02}$	—	0.55 ± 0.02
	LR (best Macro-F1)	0.57 ± 0.02	—	$\mathbf{0.56 \pm 0.02}$

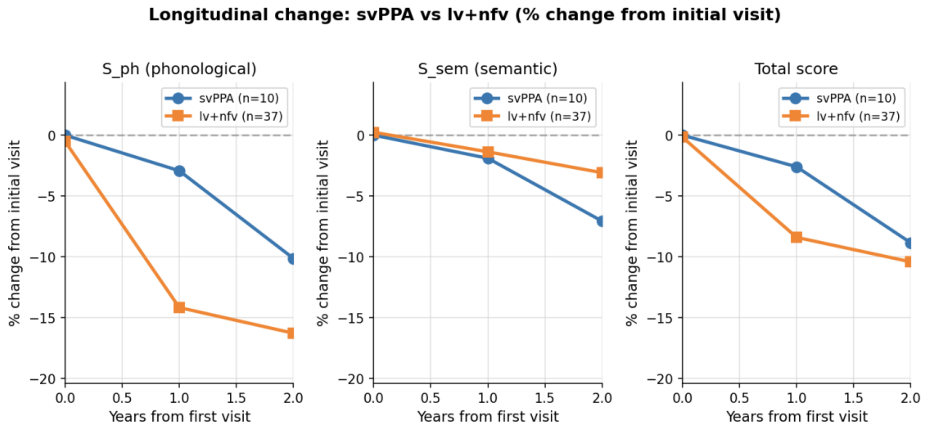

Figure 3: *Longitudinal change by subtype group, reported as percent change from first visit (years 0, 1, 2). Blue: svPPA (n=10). Orange: pooled lvPPA+nfvPPA (n=37). Panels show phonological adequacy (S_{ph}), semantic adequacy (S_{sem}), and overall score (S_{overall}). svPPA shows stronger semantic decline relative to phonological decline, while lvPPA+nfvPPA shows stronger phonological decline.*

Table 5: *LLM baseline results (qwen3.5-plus).*

Task	Acc	F1	AUC
Control vs. svPPA ($n = 50$)	0.56	0.69	0.56
svPPA vs. nfv+lvPPA ($n = 50$)	0.50	0.44	0.50

Table 6: *IPSS-PPA vs. LFTK on identical splits (mean over 10 runs). Best value per task is bolded.*

Task	Model	Acc	F1	AUC
HC vs svPPA	IPSS-PPA	0.82	–	0.92
	LFTK	**0.89**	–	**0.96**
svPPA vs nfv+lv	IPSS-PPA	**0.82**	**0.33**	**0.75**
	LFTK	0.81	0.14	0.66
Balanced 25+25	IPSS-PPA	**0.67**	**0.67**	**0.72**
	LFTK	0.57	0.56	0.62
Three-way	IPSS-PPA	0.60	**0.56**	–
	LFTK	**0.66**	0.52	–

produced a coherent, entity-rich description: *"there's also appears to be someone sitting on a pier fishing . . . a child playing in sand . . . a teenager or young person flying a kite . . . there's a house with a car in the driveway."* Scene-graph matching recovered 12 of 21 entities (recall 0.571) and 3 of 12 content units (CU coverage 0.250), yielding $S_{\text{sem}}^{\text{graph}} = 0.379$; combined $S_{\text{sem}} = 0.477$, $S_{\text{ph}} = 0.422$, $\Delta = -0.055$. Layer 2 scores were consistent with a healthy-control profile: anomia 0.219, empty speech 0.373, agrammatism 0.244, reduced speech rate 0.681. The negative Δ (phonology slightly below semantics) is atypical for svPPA and, combined with moderate but non-empty semantic content, correctly anchors the prediction to control.

Case B: misclassified control (sub-10683, ses-20181019; true: control, predicted: svPPA). The transcript contains disfluent, fragmented output with apparent code-switching artefacts and incomplete utterances: *"flag is waving, cars part in the driver . . . some people I have picked neck are under a big tree . . . the kite has a. . . "* Entity recall was 0.381 (8/21) and CU coverage fell to 0.083 (1/12), giving $S_{\text{sem}}^{\text{graph}} = 0.202$; combined $S_{\text{sem}} = 0.406$, $S_{\text{ph}} = 0.413$, $\Delta = +0.007$. Layer 2 scores were markedly elevated relative to Case A: anomia 0.649, paragrammatism 0.559, phonemic paraphasias/neologisms 0.689, agrammatism 0.501, semantic paraphasias 0.460, a profile that superficially resembles svPPA's empty-speech and semantic-poverty signature. The misclassification plausibly reflects three compounding factors: (1) poor scene-graph coverage depressed S_{sem} toward the svPPA range; (2) the near-zero dissociation ($\Delta \approx 0$) provided no phonological counter-signal; and (3) the transcript artefacts (possible ASR errors on non-English tokens) inflated Layer 2 impairment scores. This case illustrates a known failure mode of stimulus-grounded semantic scoring when ASR quality degrades: spurious tokens reduce entity-match recall without any underlying semantic impairment.

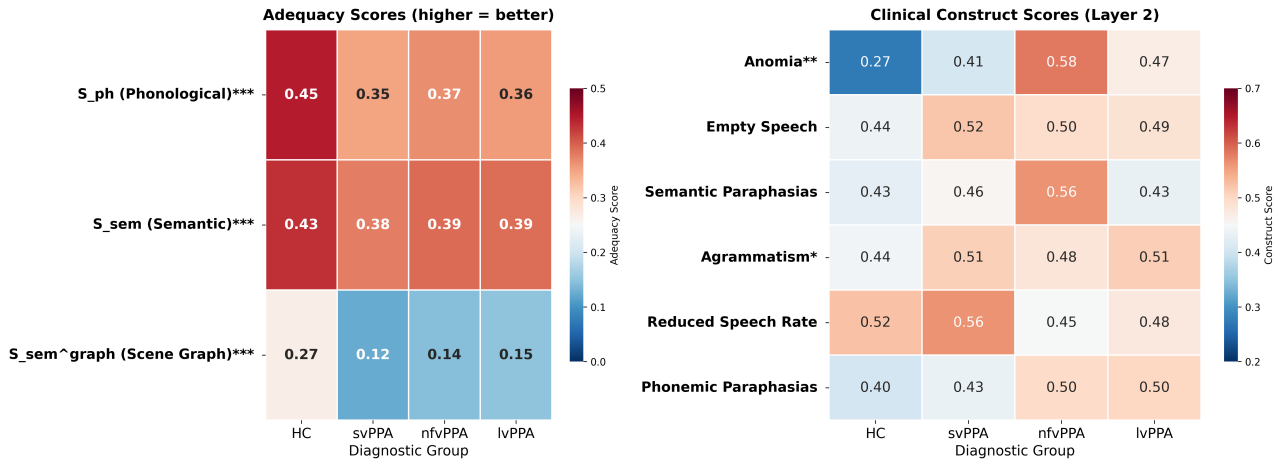

Figure 4: *Population-level adequacy and Layer 2 construct scores by diagnostic group (HC, svPPA, nfvPPA, lvPPA). Top: Mean adequacy scores (S_{ph}, S_{sem}, $S_{\text{sem}}^{\text{graph}}$); *** indicates Mann-Whitney $p < 0.001$ vs. HC. Bottom: Mean Layer 2 clinical construct scores; ** and * denote $p < 0.01$ and $p < 0.05$ respectively. Higher adequacy scores indicate more intact performance; higher construct scores indicate greater impairment on that clinical dimension.*

Table 7: *Score profiles for Cases A and B (control vs. svPPA task). Higher adequacy = more intact; higher L2 = greater impairment.*

Score	**Case A**	**Case B**
Adequacy & graph scores		
S_{ph}	0.422	0.413
S_{sem}	0.477	0.406
Δ	-0.055	$+0.007$
Entity recall	0.571	0.381
CU coverage	0.250	0.083
Proxy features (Layer 1)		
Phoneme error rate	0.578	0.587
Silence ratio	0.148	0.140
Clinical constructs (Layer 2)		
L2 reduced speech rate	0.681	0.611
L2 reduced length/complexity	0.435	0.408
L2 agrammatism	0.244	0.501
L2 paragrammatism	0.193	0.559
L2 anomia	0.219	0.649
L2 empty speech	0.373	0.289
L2 semantic paraphasias	0.289	0.460
L2 phonemic paraph./neol.	0.353	0.689
True label	control	control
Predicted label	**control**	**svPPA**

7. Discussion and Conclusion

We have presented IPSS-PPA, a fully automated pipeline for PPA subtype classification from picture description speech, integrating lexicon-free phonological scoring, scene-graph-based semantic coverage assessment, and a hierarchical layer of population-normed clinical constructs. Removing all manual preprocessing does not impair three-way subtype accuracy relative to the prior state of the art: it improves it (82.1% vs. 74.0%), suggesting that the clinically informative signal is robustly recoverable from automatically processed speech. The reduction in binary accuracy relative to Peters et al. (0.851 vs. 0.970) is consistent with the known degradation introduced by automatic diarization and with the use of a different eliciting stimulus.

A key property of the system is that every classification decision is inherently decomposable into named clinical features, without recourse to post-hoc attribution methods. Because the clinical vocabulary is constitutive of the feature space rather than imposed after model fitting, predictions can be interrogated using the same conceptual categories employed in routine assessment, providing a basis for prospective validation and integration into clinical reporting workflows. Because the Layer 2 constructs are derived from transcript-level features and the semantic graph is parameterised by an exchangeable JSON file, the pipeline generalises to new picture stimuli without retraining.

The LLM baseline confirms that structured phonological and clinical construct features capture signal that is not recoverable from transcripts alone, motivating the multimodal design. Limitations include modest F1 on the imbalanced Task 2 and the use of clinician-specified rather than learned construct weights; supervised optimisation of these quantities and replication on an independent cohort are priorities for future work. More broadly, the present work demonstrates that a compact, interpretable feature set grounded in established clinical theory can match or exceed high-dimensional systems on PPA subtyping, a finding with implications for the design of automated speech biomarker pipelines more generally, where clinical deployability and transparency are as important as classification accuracy.

8. Acknowledgments

The authors thank the participants and clinical staff of the UCSF Memory and Aging Center for their contributions to the longitudinal PPA cohort. We are especially grateful to Lynn Kurteff and the UCSF clinical team (Zoe Ezzes, Willa Keegan-Rodewald, Jet M.J. Vonk, Siddarth Ramkrishnan, Giada Antonicelli, Zachary A. Miller, and Maria Luisa Gorno-Tempini) for their clinical expertise and iterative guidance in grounding the construct framework in established diagnostic practice, and for their stewardship of the longitudinal cohort that made this work possible. We also thank Emma Yang for contributions to scene graph annotation and entity mapping. Data access coordination was supported by the UCSF MAC research team.

This work was supported in part by [grant information to be added in camera-ready version].

In accordance with ISCA policy, the authors disclose that a large language model assistant was used for only editing and polishing the prose of this manuscript, specifically for improving sentence clarity, correcting grammatical errors, and improving the consistency of technical terminology. All scientific claims, design decisions, experimental results, and interpretations are the sole responsibility of the human authors, who accept full accountability for the work.

9. Generative AI Use Disclosure

In accordance with ISCA policy, the authors disclose that a large language model (LLM) assistant was used exclusively for editing and polishing the prose of this manuscript, specifically for improving sentence clarity, correcting grammatical errors, and improving the consistency of technical terminology across sections. The LLM was not used to generate any scientific content, including experimental design, feature engineering, model architecture, data analysis, result interpretation, or any figures or tables. All such content is the original work of the human authors. All co-authors have reviewed the final manuscript, accept full responsibility for its content, and consent to its submission.

10. References

[1] M. M. Mesulam, "Primary progressive aphasia and the language network: the 2013 H. Houston Merritt lecture," *Neurology*, vol. 81, no. 5, pp. 456–462, 2014.

[2] M. L. Gorno-Tempini, A. E. Hillis, S. Weintraub, A. Kertesz, M. Mendez, S. F. Cappa, J. M. Ogar, J. D. Rohrer, S. Black, B. F. Boeve *et al.*, "Classification of primary progressive aphasia and its variants," *Neurology*, vol. 76, no. 11, pp. 1006–1014, 2011.

[3] G. Hickok and D. Poeppel, "The cortical organization of speech processing," *Nature Reviews Neuroscience*, vol. 8, no. 5, pp. 393–402, 2007.

[4] N. Nevler, S. Ash, C. Jester, D. J. Irwin, M. Liberman, and M. Grossman, "Validated automatic speech biomarkers in primary progressive aphasia," *Annals of Clinical and Translational Neurology*, vol. 6, no. 8, pp. 1518–1529, 2019.

[5] C. Themistocleous, M. Eckerström, and D. Kokkinakis, "Identification of PPA and its variants using machine learning," in *Proceedings of the 9th International Conference on Language Resources and Evaluation*, 2018, pp. 3399–3404.

[6] K. C. Fraser, J. A. Meltzer, N. L. Graham, C. Leonard, G. Hirst, S. E. Black, and E. Rochon, "Automated classification of primary progressive aphasia subtypes from narrative speech transcripts," *Cortex*, vol. 55, pp. 43–60, 2014.

[7] C. Themistocleous, B. Ficek, K. Webster, D.-B. den Ouden, A. E. Hillis, and K. Tsapkini, "Automatic subtyping of individuals with primary progressive aphasia," *Journal of Alzheimer's Disease*, vol. 79, no. 3, pp. 1185–1194, 2021.

[8] N. Rezaii, D. Hochberg, M. Quimby, B. Wong, M. Brickhouse, A. Touroutoglou, B. C. Dickerson, and P. Wolff, "Artificial intelligence classifies primary progressive aphasia from connected speech," *Brain*, vol. 147, no. 9, pp. 3070–3082, 2024.

[9] F. Peters, W. R. Bevan-Jones, G. Threlfall, J. M. Harris, J. S. Snowden, M. Jones, J. C. Thompson, D. J. Blackburn, and H. Christensen, "Automatic detection and sub-typing of primary progressive aphasia from speech: Integrating task-specific features and spatio-semantic graphs," in *Proceedings of Interspeech*, 2025.

[10] J. M. J. Vonk, J. Lian, Z. Ezzes, C. J. Cho, B. T. Morin, R. Bogley, Z. Miller, M. L. Mandelli, G. Anumanchipalli, and M. L. Gorno-Tempini, "Automated lexical dysfluency analysis to differentiate primary progressive aphasia variants," in *Alzheimer's Association International Conference*, 2025.

[11] J. Lian, C. Feng, N. Farooqi, S. Li, A. Kashyap, C. J. Cho, P. Wu, R. Netzorg, T. Li, and G. K. Anumanchipalli, "Unconstrained dysfluency modeling for dysfluent speech transcription and detection," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[12] L. E. Nicholas and R. H. Brookshire, "A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 2, pp. 338–350, 1993.

[13] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "Pyannote.audio: neural building blocks for speaker diarization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 8210–8214.

[14] V. C. Zimmerer, M. Wibrow, and R. A. Varley, "Formulaic language in people with probable Alzheimer's disease: a frequency-based approach," *Journal of Alzheimer's Disease*, vol. 53, no. 3, pp. 1145–1160, 2016.

[15] H. Phan *et al.*, "Automated picture description assessment with task-grounded visual context," 2024, preprint.

[16] A. Favaro, M. Cao, T. Gessesse, S. Ghosh, A. Luzardo, L. Moro-Velazquez, A. Butala, J. Villalba, and N. Dehak, "Automatic detection of Alzheimer's disease from speech and natural language: a systematic literature review," *Frontiers in Aging Neuroscience*, vol. 15, p. 1210894, 2023.

[17] D. Ambadi, J. R. Hodges, M. Irish, S. Ahmed, and O. Piguet, "A spatio-semantic framework for characterizing picture description in neurodegenerative disorders," *Frontiers in Human Neuroscience*, vol. 15, p. 626780, 2021.

[18] J. Lian and G. Anumanchipalli, "Towards hierarchical spoken language disfluency modeling," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Y. Graham and M. Purver, Eds. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 539–551. [Online]. Available: https://aclanthology.org/2024.eacl-long.32

[19] J. Lian, X. Zhou, Z. Ezzes, J. Vonk, B. Morin, D. P. Baquirin, Z. Miller, M. L. Gorno Tempini, and G. Anumanchipalli, "Ssdm: Scalable speech dysfluency modeling," in *Advances in Neural Information Processing Systems*, vol. 37, 2024.

[20] J. Lian, X. Zhou, C. Guo, Z. Ye, Z. Ezzes, J. Vonk, B. Morin, D. Baquirin, Z. Mille, M. L. G. Tempini, and G. K. Anumanchipalli, "Automatic detection of articulatory-based disfluencies in primary progressive aphasia," *IEEE JSTSP*, 2025.

[21] X. Zhou, A. Kashyap, S. Li, A. Sharma, B. Morin, D. Baquirin, J. Vonk, Z. Ezzes, Z. Miller, M. Tempini, J. Lian, and G. Anumanchipalli, "Yolo-stutter: End-to-end region-wise speech dysfluency detection," in *Interspeech 2024*, 2024, pp. 937–941.

[22] J. Zhang, X. Zhou, J. Lian, S. Li, W. Li, Z. Ezzes, R. Bogley, L. Wauters, Z. Miller, J. Vonk *et al.*, "Analysis and evaluation of synthetic data generation in speech dysfluency detection," *Interspeech*, 2025.

[23] C. Guo, J. Lian, X. Zhou *et al.*, "Dysfluent wfst: A framework for zero-shot speech dysfluency transcription and detection," *Interspeech*, 2025.

[24] C. Guo, J. Lian, Y. Liu, B. Huang, S. Narayanan, C. J. Cho, and G. Anumanchipalli, "Huper: A human-inspired framework for phonetic perception," *arXiv preprint arXiv:2602.01634*, 2026.

[25] Z. Ye, J. Lian, X. Zhou, J. Zhang, H. Li, S. Li, C. Guo, A. Das, P. Park, Z. Ezzes *et al.*, "Seamless dysfluent speech text alignment for disordered speech analysis," *Interspeech*, 2025.

[26] J. M. Vonk, J. Lian, C. J. Cho, G. Antonicelli, Z. Ezzes, L. D. Wauters, W. Keegan-Rodewald, G. L. Kurteff, D. A. Rodriguez, N. Dronkers *et al.*, "Ai-based speech error detection to differentiate primary progressive aphasia variants," *medRxiv*, pp. 2026–02, 2026.

[27] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the International Conference on Machine Learning*, 2023, pp. 28 492–28 518.

[28] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spacy: Industrial-strength natural language processing in python," 2020.

[29] S. M. Wilson, D. K. Eriksson, S. M. Schneck, and J. M. Lucanie, "A quick aphasia battery for efficient, reliable, and multidimensional assessment of language function," *PLOS ONE*, vol. 13, no. 2, p. e0192773, 2018.

[30] Carnegie Mellon University Speech Group, "The cmu pronouncing dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict, 2014, version 0.7b, accessed 2026-03-04.